DOCUMENT RESUME

ED 309 187 TM 013 658

AUTHOR van der Linden, Wim J.; Adema, Jos J.
TITLE Algorithmic Test Design Using Classical Item Parameters. Project Psychometric Aspects of Item Banking No. 29. Research Report 88-2.
INSTITUTION Twente Univ., Enschede (Netherlands). Dept. of Education.
PUB DATE Mar 88
NOTE 37p.
AVAILABLE FROM Bibliotheek, Department of Education, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Algorithms; Computer Simulation; Estimation (Mathematics); Foreign Countries; *Item Banks; *Latent Trait Theory; Linear Programing; Mathematical Models; *Test Construction; Test Theory
IDENTIFIERS Alpha Coefficient; Classical Test Theory; Empirical Methods; *Item Parameters; *Rasch Model; Three Parameter Model

ABSTRACT

Two optimalization models for the construction of tests with a maximal value of coefficient alpha are given. Both models have a linear form and can be solved by using a branch-and-bound algorithm. The first model assumes an item bank calibrated under the Rasch model and can be used, for instance, when classical test theory has to serve as an interface between the item bank system and a user not familiar with modern test theory. Maximization of alpha was obtained by inserting a special constraint in a linear programming model. The second model has wider applicability and can be used with any item bank for which estimates of the classical item parameter are available. The models can be expanded to meet practical constraints with respect to test composition. An empirical study with simulated data using two item banks of 500 items was carried out to evaluate the model assumptions. For Item Bank 1 the underlying response was the Rasch model, and for Item Bank 2 the underlying model was the three-parameter model. An appendix discusses the relation between item response theory and classical parameter values and adds the case of a multidimensional item bank. Three tables present the simulation study data. (SLD)

**********
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
**********

ED309187

# Algorithmic Test Design Using Classical Item Parameters

Research Report 88-2

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
J. NELISSEN
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Wim J. van der Linden
Jos J. Adema

department of EDUCATION

University of Twente

Division of Educational Measurement and Data Analysis

TM013668

Project Psychometric Aspects of Item Banking No.29

Colofon:
Typing: Mevr. L.A.M. Padberg
Cover design: Audiovisuele Sectie TOLAB Toegepaste Onderwijskunde
Printed by: Centrale Reproductie-afdeling

# Algorithmic Test Design
## Using Classical Item Parameters

Wim J. van der Linden

Jos J. Adema

Algorithmic Test Design Using Classical Item Parameters / Wim J. van der Linden & Jos J. Adema – Enschede : University of Twente, Department of Education, March, 1988. – 30 p.

## Abstract

Two optimization models for the construction of tests with a maximal value of coefficient alpha are given. Both models have a linear form and can be solved using a branch-and-bound algorithm. The first model assumes an item bank calibrated under the Rasch model and can be used, for instance, when classical test theory has to serve as an interface between the item bank system and a user not familiar with modern test theory. The second model has wider applicability and can be used with any item bank for which estimates of the classical item parameters are available. The models can be expanded to meet practical constraints with respect to test composition. An empirical study with simulated data was carried out to evaluate the model assumptions.

# Algorithmic Test Design
# Using Classical Item Parameters

A useful phenomenon in educational and psychological measurement is the construction of customized tests from item banks. An item bank is a large collection of test items all measuring the same ability or domain of knowledge, stored in a computer together with empirical estimates of their properties. The fact that the item properties are known allows the test constructor to have explicit control of them and select optimal tests.

If the properties of the items are modeled using an item response theory (IRT) model, estimates of parameters representing such properties as item difficulty, discriminating power, and the effect of random guessing are stored in the item bank. Although item selection can be based on these parameter values, a more advanced procedure uses the item and test information function from IRT. Birnbaum (1968) and Lord (1980) suggested a procedure in which the test constructor first specifies a target for the test information function and then selects the items such that the sum of their information functions meets the target. Theunissen (1985) presented a zero-one programming model for selecting a test of minimal length subject to the condition that its information function is not below the target. The model can be solved using one of the branch-and-bound algorithms available in the literature (e.g., Wagner, 1975). Alternative models and procedures have been given by Adema (1988),

Boekkooi-Timminga (1987), Boekkooi-Timminga and van der Linden (1987), Theunissen (1986), Theunissen and Verstralen (1986), van der Linden (1987), and van der Linden and Boekkooi-Timminga (1988a, 1988b).

Although item banking and IRT are natural partners (van der Linden, 1986a), this does not necessarily imply that test construction has to be based solely on information functions. The following two examples refer to practical cases in which item selection based on parameters from classical test theory may be helpful:

1. The item bank has been calibrated under an IRT model, but some of the users are not familiar with the theory and want to have the option of using classical item parameters. In such cases, it is possible to use the classical test theory as an interface between the item bank system and its users. The system then predicts the classical parameter values for the population of examinees concerned (see the Appendix) enabling the users to select tests with optimal values for the parameters.
2. The examinees are sampled from a population with a fixed ability distribution. Therefore, for certain applications the use of classical item parameters may be feasible. For example, the classical index $\pi$ roughly orders the difficulties of the items for a random examinee, and the availability of estimates of the item $\pi$ values is considered as sufficient to base test construction on. Item banking with classical item

parameters is dealt with extensively in de Gruijter (1986).

The present paper was motivated by an item banking project in which the need of the option in the former example was felt. The first test construction model presented below deals with this case. The second model is more general and also has applicability in other cases where item selection is based on classical parameters. Both models are zero-one programming models that maximize (a linearized version of) the well-known coefficient alpha. Results from an empirical study with simulated data to verify the model assumptions follow the presentation of the models.

## Maximal Test Reliability as a Classical Goal

A classical goal in test construction is maximization of the reliability of the test for a given population of examinees. Since the reliability coefficient can only be estimated from hard-to-realize replicated measurements, in practice coefficient alpha, a well-known and simple lower bound to the test reliability, is mostly used.

Let $\sigma_i^2$ denote the variance of the scores on item i for the given population, and let $\rho_{iX}$ represent the item-test correlation. For a test of n items, coefficient alpha is defined as:

$$(1) \qquad \alpha \equiv n(n-1)^{-1}[1 - (\sum_{i=1}^{n} \sigma_i^2)\sigma_X^{-2}]$$

(Lord & Novick, 1968, sects. 4.4). Since

$$(2) \qquad \sigma_X^2 = (\sum_{i=1}^{n} \sigma_i \rho_{iX})^2$$

(Lord & Novick, 1986, sect. 15.3), the right-hand term in the bracketed expression is equal to

$$(3) \qquad (\sum_{i=1}^{n} \sigma_i^2)(\sum_{i=1}^{n} \sigma_i \rho_{iX})^{-2} \quad .$$

For a test of fixed length, maximization of alpha is equivalent to minimization of (3).

A zero-one programming model for maximization of alpha can now be formulated as follows. For each item i = 1, ..., I the decision variable $x_i$ is defined:

$$(4) \qquad x_i = \begin{cases} 0 & \text{if item i is not in the test} \\ 1 & \text{if item i is in the test.} \end{cases}$$

A maximal value of alpha is obtained for a solution to the following problem:

$$(5) \qquad \text{minimize } (\sum_{i=1}^{I} \sigma_i^2 x_i)(\sum_{i=1}^{I} \sigma_i \rho_{iX} x_i)^{-2} .$$

subject to

$$(6) \qquad \sum_{i=1}^{I} x_i = n.$$

(7) $x_i \in \{0, 1\}$ , $i = 1, \ldots, I.$

Although the model is of the zero-one type, it has a nonlinear objective function. Efficient algorithms for solving such models are not known (Garfinkel & Nemhauser, 1972). In addition, a minor problem in (5) is the dependency of $\rho_{iX}$ on the unknown test score. As is usual in classical test construction, this problem will be ignored. Also, for an item bank system with an underlying IRT model, it is easy to predict the correlation between the item score and the number-right score for the complete bank. This constant could be substituted for $\rho_{iX}$ in (5).

Two alternative linear models will be given for which practical algorithms do exist. In the first model, a condition for alpha to be maximal is inserted as a linear constraint into the model. The condition can be shown to apply for an item bank calibrated under the Rasch (1980) model. The second model does not assume any IRT model. In this model, a linearized part of (3) is used as objective function, whereas the remaining part serves as a linear constraint. The two models will now be derived.

## Maximal Alpha as a Linear Constraint

For the two-parameter normal-ogive model, a simple relation between the discrimination parameter and the item-ability correlation exists. Also, it is known that the logistic

function approximates the normal ogive excellently. On these two findings the following derivation of a sufficient condition for alpha to be maximal is based.

Derivation of the Condition

For an ability $\theta$, the normal-ogive model is defined as:

$$(8) \qquad p_i(\theta) = \int_{-\infty}^{a_i(\theta-b_i)} (2\pi)^{-1}\exp(-u^2/2)du \ ,$$

where $a_i$ and $b_i$ are the parameters for the discriminating power and difficulty of item i, and $p_i(\theta)$ is the probability of a correct response on i for an examinee with ability $\theta$.

If latent response variables $Y_i$, $i = 1, \ldots, I$, are assumed such that $Y_i \geq y_i$ generates a correct response, but $Y_i < y_i$ an incorrect one, and the distributions of $Y_i$ given $\theta$ are normal with linear regression functions and homoscedasticity, the following relation exists between $a_i$ and the item-ability (biserial) correlation $\rho_{i\theta}$:

$$(9) \qquad \rho_{i\theta} = a_i(1+a_i^2)^{-1/2}$$

(Lord & Novick, 1968, sects. 16.8 – 16.10). Since, for a scale factor 1/1.7 in (8), the logistic and normal-ogive curves are known to approximate each other by less than 0.01 uniformly in $\theta$ (Haley, in Lord & Novick, 1968, sect. 17.2; for improvements on this well-known result, see Molenaar,

1974) and the Rasch model is a logistic model with $a_i = 1$ for all items, it follows that in the Rasch model $\rho_{i\theta}$ is approximately constant. Hence, if for all items $\rho_{iX}$ has the same relation to $\rho_{i\theta}$, it is also a constant. In this case (2) reduces to

$$(10) \qquad \sigma_X^2 = c(\sum_{i=1}^{n} \sigma_i)^2 \ ,$$

with $c \geq 0$. It follows that

$$(11) \qquad \sigma_X^2 = c(\sum_{i=1}^{n} \sigma_i^2 + \sum_{i \neq j} \sigma_i \sigma_j)$$

Substituting this result into (3), yields

$$(12) \qquad c^{-1}[1 + (\sum_{i \neq j} \sigma_i \sigma_j)(\sum_{i=1}^{n} \sigma_i^2)^{-1}]^{-1} \ .$$

Observe that (12) now is invariant under multiplication of $(\sigma_1, \ldots, \sigma_n)$ by a constant. Without loss of generality, $\sum_{i=1}^{n} \sigma_i^2$ can therefore be taken to be equal to a constant $k > 0$. This shows that minimization of (12) amounts to maximization of $\sum_{i \neq j} \sigma_i \sigma_j$. However, (11) implies that this is also equivalent to maximization of $\sigma_X^2$. Hence, it follows from (10) that (3) has a minimum for the value of $(\sigma_1, \ldots, \sigma_n)$ maximizing

$$(13) \qquad \sum_{i=1}^{n} \sigma_i$$

under the condition that

$$(14) \qquad \sum_{i=1}^{n} \sigma_i^2 = k.$$

Maximization using Lagrange multipliers results in the following system of equations:

$$(15) \qquad 1 + 2\lambda\sigma_i = 0, \qquad i = 1, \ldots, n;$$

$$(16) \qquad \sum_{i=1}^{n} \sigma_i^2 - k = 0, \qquad i = 1, \ldots, n.$$

Since k is arbitrary, the system is solved for

$$(17) \qquad 0 < \sigma_1 = \ldots = \sigma_n.$$

Thus, provided the assumptions leading to (10) are satisfied, coefficient alpha is maximal if and only if

$$(18) \qquad 0 < \chi_1 = \ldots = \chi_n < 1,$$

where $\chi_i = \pi_i$ or $1 - \pi_i$, and $\pi_i$ is the classical difficulty parameter for item i. Without loss of generality, in the following only the case of equal $\pi_i$ values will be considered.

Dependent on the composition of the item bank, the solution need not be unique and simultaneous optimization with respect to another goal can be possible. This result is used in the following linear programming (LP) model.

LP Model

It is assumed that the estimates of the item $\pi$ values are rounded to a significant digit such that larger classes of items with the same rounded value exist. The sets of indices of items in the same class are denoted as $I_1, \ldots, I_j, \ldots, I_J$. As an example of simultaneous optimization with respect to a second goal, it is assumed that realistic estimates of the time needed to solve the items in the bank exist and that the goal is to minimize the total administration time needed for the test. Let $t_i$ be such an estimate for item i, e.g., an estimate of the 95th percentile in the distribution of time needed to solve item i for the given population.

The following linear model realizes (18) at the same time minimizing the total administration time of the test:

$$\text{(19)} \qquad \text{minimize } \sum_{i=1}^{I} t_i x_i$$

subject to

$$\text{(20)} \qquad \sum_{i \in I_j} x_i - n y_j = 0, \qquad j = 1, \ldots, J,$$

$$(21) \qquad \sum_{i=1}^{I} x_i = n,$$

$$(22) \qquad x_i,\ y_j \in \{0,\ 1\}, \qquad i = 1,\ \ldots,\ I,\quad j = 1,\ \ldots,\ J.$$

The additional decision variable $y_j$ in the model indicates from which class the items are selected. The constraints in (21) guarantees that exactly n items are selected from the I items in the bank. The constraints in (20) and (21) together allow $y_j$ to take the value one exactly once. The model in (19) through (22) is linear and can be solved by a standard branch-and-bound algorithm from the operations research literature. Adema (1988) gives a modified branch-and-bound procedure that reduces the CPU-time needed for a standard procedure considerably.

The above model is too simple to deal with most test construction problems. In practice, usually various kinds of restrictions with respect to, e.g., item content, simultaneous inclusion of different items, or ranges of possible item-parameter values exist. This point will be taken up after the presentation of an alternative model.

## A Linearized Version of Alpha as Objective Function

In the previous model, a maximal value of alpha was realized by adopting a special constraint in the model. The following model explicitly maximizes alpha by a direct attack of (3).

Inspection of (5) shows that both of its sums are linear in the decision variables. This suggests an approach in which one of these expressions is used as objective function and the other as a constraint. Since for a wide range of possible values of $\pi$, the numerator of (3) varies less than the denominator, alpha can be expected to depend stronger on the latter. This effect is verified empirically in Ebel (1967). Therefore, it seems sensible to maximize the denominator of (3) constraining the numerator to a low value. This is realized in the following model

$$\text{(23)} \qquad \text{maximize } \sum_{i=1}^{I} \sigma_i \rho_{iX} x_i$$

subject to

$$\text{(24)} \qquad \sum_{i=1}^{I} \sigma_i^2 x_i \leq v,$$

$$\text{(25)} \qquad \sum_{i=1}^{I} x_i = n,$$

$$\text{(26)} \qquad x_i \in \{0, 1\}, \qquad i = 1, \ldots, I,$$

where $v > 0$ is a constant. Again, the model is linear and can be solved for $(x_1, \ldots, x_I)$ by one of the branch-and-bound algorithms referred to earlier.

The choice of a value for v can be motivated as follows.

The maximal possible value of $\sum_{i=1}^{I} \sigma_i^2$ in the model is equal to n/4. In addition, the numerator and denominator of (3) have $\sigma_i$ as a common factor. Therefore, if v approaches its maximum, a maximal value will be found, but at the same time the numerator will tend to be too large. On the other hand, if v approaches its minimum, a minimal value for the numerator will be attained but at the cost of a constrained denominator. Now this is due not only to the common factor $\sigma_i$, but also to a restriction-of-range effect on $\rho_{iX}$. Hence, the optimal value of v will tend to be closer to n/4 than to zero. This issue will be pursued further in the section on empirical results below. It should be noted that by varying v all possible tests of length n can be produced as a solution to the model. So in principle the structure of the model does not preclude any possible test from showing up as optimal.

## Possible Additional Constraints

As already noted, in order to solve most practical test construction problems, the above models have to be made more realistic. For example, a test constructor may want to have control of such features of the test as its validity with respect to several domains of content represented in the item bank. This and all other possible demands can be adopted in the above models, provided they are formulated as linear constraints. The models can then still be solved by the same class of algorithms and the solution always automatically

meets the new constraints. An extensive review of possible constraints from the practice of test construction that can be formulated in a linear form is given in van der Linden and Boekkooi-Timminga (1988b). The review includes constraints controlling the composition of the test with respect to behavioral dimensions and item content and format; item parameters like administration time, frequency of previous administrations and item difficulty; curriculum differences between groups; inclusion or exclusion of individual items; and dependencies between the items. The following example illustrates some of the possibilities.

Example

A test with maximal value of coefficient alpha has to be constructed from a Physics item bank. From each of the topics $p = 1, \ldots, P$ covered by sets of items, $V_p$, in the bank, the test constructor wants $n_p$ items in the test. The items have also been classified with respect to a behavioral dimension (e.g., knowledge of facts, concepts, application of rules) and from each of the sets $V_q$, $q = 1, \ldots, Q$, at least $n_q$ items should be in the test. The estimated time in minutes needed to solve the items in the bank, $t_i$, $i = 1, \ldots, I$, (see above) is known and the total administration time is not allowed to exceed T minutes. Also, for each item it has been recorded how often it was administered before, and only items with a frequency of previous administration, $f_i$, not larger than one are allowed in the test. Items with a multiple-choice format, collected in subset $V_s$, should be excluded

from the test. Finally, for some special reason the test constructor wants item #115 in the test, and items #19 and #203 may not be chosen together. All these demands are realized in the following model:

(27) maximize $\sum_{i=1}^{I} \sigma_i \rho_{iX} x_i$

subject to

(28) $\sum_{i=1}^{I} \sigma_i^2 x_i \leq v,$

(29) $\sum_{i \in V_p} x_i = n_p,$ $p = 1, \ldots, P,$

(30) $\sum_{i \in V_q} x_i \geq n_q,$ $q = 1, \ldots, Q,$

(31) $\sum_{i=1}^{I} t_i x_i \leq T,$

(32) $f_i x_i \leq 1,$ $i = 1, \ldots, I,$

(33) $\sum_{i \in V_s} x_i = 0,$

(34) $x_{115} = 1.$

$$x_{19} + x_{203} \leq 1. \tag{35}$$

$$x_i \in \{0, 1\}, \qquad i = 1, \ldots, I. \tag{36}$$

It should be noted that when specifying the constants in the model, certain relations ought to be obeyed. For instance, no feasible solution exists if $\sum_{q=1}^{Q} n_q > \sum_{p=1}^{P} n_p$. Further, constraints (32) – (34) do not enter the actual optimization procedure; they only reduce the number of decision variables.

## Empirical Validation of Model Assumptions

Two item banks of 500 items were generated to evaluate the model assumptions. For Item Bank 1, the underlying response model was the Rasch model with item parameters drawn from the distribution N(–0.5, 1). For Item Bank 2, the underlying model was the 3–parameter model with item parameters $a_i$ and $b_i$ drawn from the distributions U(0.5, 1.5) and U(–3, 1), respectively. The guessing parameter $c_i$ was set equal to 0.1. To estimate the item difficulties, $p_i$, and item discriminations (i.e., item–test correlations, where the whole item bank is considered as the test), $r_{ib}$, 1,000 examinees ($\theta \sim N(0,1)$) were generated to answer the items.

The program Lando (Anthonisse, 1984) was used to solve the zero–one programming models on a DEC 2060 computer. Because it takes too much time to find a zero–one solution for the model in (23) to (26), the relaxation of this model

was solved, i.e., the model with the constraints $0 \leq x_i \leq 1$, $i = 1, 2, \ldots, I$ instead of $x_i \in \{0, 1\}$. This could be done, because it is known (Dantzig, 1957) that the number of fractional values in the solution is not greater than the number of constraints. Therefore, the solution to the model in (23) to (26) was found by rounding at most two fractional values.

The model assumptions were first verified by comparing tests from Item Bank 1 for different values of v and p. The number of items in the tests was 20. Table 1 shows the values of coefficient $\alpha$. In this table, $\alpha^*$ denotes coefficient alpha with item-bank correlations replacing item-test correlations, whereas $\alpha$ is the exact value of the coefficient calculated after the test was selected.

---

Insert Table 1 here

---

Table 1 shows that the differences between values of $\alpha$ of tests constructed for different values of v were small. Higher values of v gave the best results. The values of $\alpha$ for the model with maximal alpha as a constraint were not as good as for the other model. From Table 1 it is also clear that the results were worse, the greater the difference between the chosen value of p and .5. The same trend was observed in extensive simulations not reported here (see Adema, 1987).

Apparently, the assumptions leading to the well-known result in (9) or the assumption of $\rho_{iX}$ having the same relation to $\rho_{i\theta}$ for all items, are not entirely met for arbitrary data sets.

For Item Bank 2 (3-parameter model), only the model with a linearized version for $\alpha$ as objective function was applicable. Again, tests were constructed for different values of v. The results are displayed in Table 2.

---

Insert Table 2 here

---

Once more the best results were found for high values of v. Therefore, it is possible to choose v maximal so that constraint (24) is redundant and can be omitted.

Because the variances of the items are not as important as the item discriminations, the following zero-one programming model was also tried out:

$$(37) \qquad \text{maximize} \sum_{i=1}^{I} \rho_{iX} x_i$$

subject to

$$(38) \qquad \sum_{i=1}^{I} x_i = n,$$

(39) $x_i \in \{0, 1\}$, $i = 1, 2, \ldots, I$.

In Table 3, values of $\alpha$ are shown for tests constructed with model (23), (25), and (26) (v maximal) and with model (37) through (39). The number of items in the tests was 20 or 40 and the models were applied to both item banks.

---

Insert Table 3 here

---

Table 3 demonstrates that model (37) to (39) gave very good results. The values of $\alpha$ were as good as for the best choices of v in Table 1 and 2.

Table 1, 2, and 3 show that it is possible to construct tests with item-test correlations replaced by item-bank correlations, because generally tests with a high value for $\alpha^*$ also have a high value for $\alpha$.

## Discussion

Two models for maximization of coefficient alpha as a function of classical item parameters were presented.

In the first model maximization of alpha was obtained by inserting a special constraint in a linear programming model. The fact that minimization of administration time was chosen as explicit objective function was just for the purpose of

illustration. Measures for other aspects, e.g., for curricular fit of the test or uniform usage of the items in the bank (van der Linden & Boekkooi-Timminga, in press), could also have been optimized. The important point to note is that the model allows optimization with respect to two different objectives. The model is based on a formalization of the intuitive notion that an item bank conforming to the Rasch model should consists of items with equal (classical) discriminating power. However, the formalization, which resulted in (18) as a condition for alpha to be maximal, also needed extra assumptions in addition to the Rasch model. As shown in Table 1, for items satisfying the condition in (18) $\alpha$ tends to decrease if $\pi_i$ deviates from .50. For $.30 < \pi_i < .70$ the results are still satisfactory but outside this interval $\alpha$ drops relatively quickly. Since the data were generated under the Rasch model, this phenomenon implies that the extra assumptions are not tenable for all possible data sets. Therefore, models as in (19) to (22) are only recommended for items with values for $\pi_i$ in this interval.

The second model is universal in the sense that it does not assume any IRT model or other assumptions about the items. The model is a direct attack of the kernel of alpha in (3): it maximizes the denominator at the same time constraining the numerator. Ample experience with the model for various types of data has shown that the solution invariably produces the maximal value for alpha for v close to n/4 (maximum of $\Sigma_{i=1}^{I} \sigma_i^2$ in the model). For example, for n = 500, all simulations produced the maximum of alpha

for v in the neighborhood of 95% of n/4. Also, the optimal value of alpha increases monotonically with v to the point at which the maximum is obtained and then shows a monotonic but slight decrease. Therefore, for large item banks, n > 500, say, it is recommended to set v at its maximal value. However, as already observed, the model in (37) through (39) almost always produced comparable results. If no additional constraints have to be met, this model can be solved by a simple algorithm that picks items with the largest values for their item-test correlations. For such applications, the model is strongly recommended.

Finally, it is observed that the advantage of a linear programming approach to test construction lies not only in its power to optimize a test parameter as coefficient alpha, but also in the possibility to include additional practical constraints. The example given earlier shows that the presence of such constraints easily involves combinatorial problems that cannot be solved by hand.

## APPENDIX

The availability of a item bank system with items calibrated under an IRT model allows the possibility to use classical test theory as an interface between the system and a user not familiar with IRT. For a given population of examinees the system is able to predict the values of the classical parameters for the items. These values can be used as the input of one of the models in the paper, whereafter the

system predicts the values for the test parameters of the resulting test.

Jensema (1976) and Lord (1980) deal with the relation between IRT and classical parameter values. The following summarizes some of the results and adds the case of a multidimensional item bank. A complete treatment is given in van der Linden (1986b).

Let $p_i(\theta)$ be the probability of a correct response on item i for an examinee with ability $\theta$ explained by the IRT model and let $F(\theta)$ be the distribution function for the population of examinees under consideration. The basic equations are:

$$(1) \quad \pi_i = \int_{-\infty}^{+\infty} p_i(\theta) dF(\theta) \ .$$

$$(2) \quad \pi_{ij} = \int_{-\infty}^{+\infty} p_i(\theta) p_j(\theta) \ dF(\theta) \ .$$

The first equation gives the classical item difficulty; the second equation uses the property of local independence and is necessary to derive the item-test correlation:

$$(3) \quad \rho_{iX} = \sigma_{iX} \ (\sigma_i \sigma_X)^{-1} \ .$$

This follows from

$$(4) \quad \sigma_i = \pi_i(1-\pi_i) \ .$$

(5) $\sigma_{iX} = \Sigma_j(\pi_{ij}-\pi_i\pi_j)$ .

(6) $\sigma_X^2 = \Sigma_i\pi_i(1-\pi_i) + \Sigma_{i\neq j}(\pi_{ij}-\pi_i\pi_j)$ .

Case of Multidimensionality

Suppose the item bank falls apart into two different sets of items and that for each set an IRT model holds. Let $\theta_v$ and $\theta_w$ be the ability parameters spanning the items in each set, while $F_v(\theta_v)$, $F_w(\theta_w)$, and $F_{vw}(\theta_v,\theta_w)$ are now the distribution functions for the given population of examinees. If $i_v$ and $i_w$ denote an arbitrary item in the two respective sets, then the basic equations are

(7) $\pi i_v = \int_{-\infty}^{+\infty} p_{i_v}(\theta_v)dF_v(\theta_v)$ .

(8) $\pi i_v i_w = \int_{-\infty}^{+\infty} p_{i_v}(\theta_v)\, p_{i_w}(\theta_w)dF_{vw}(\theta_v,\theta_w)$ .

Equation (8) assumes the property of local independence for response variables associated with items from different sets. The property can be proven to hold as follows:

Proof. Let $(U_{i_v})$ and $(U_{i_w})$ be the response variables associated with the two sets of items. Since for $(U_{i_v})$ an IRT model holds, $\theta_v$ spans this set completely. Thus, no partition of the population of examinees is possible that introduces different distributions over $(U_{i_v})$ for a given value of $\theta_v$. Therefore, the values of $(U_{i_w})$ cannot introduce such a partition and the variables in $(U_{iv})$ are locally independent of those in $(U_{iw})$. []

## References

Adema, J.J. (1987). Toetsconstructie met klassieke item en test parameters [Test construction using classical item and test parameters] (Rapport 87-1). Enschede: University of Twente, Department of Education.

Adema, J.J. (1988). A note on solving large-scale zero-one programming problems (Research Report 88-4). Enschede: University of Twente, Department of Education.

Anthonisse, J.M. (1984). Lando. Amsterdam: Centre for Mathematics and Computer Science CWS.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F.M. Lord & M.R. Novick, Statistical theories of mental test scores. Reading, Mass.: Addison-Wesley.

Boekkooi-Timminga, E. (1987). Simultaneous test construction by zero-one programming. Methodika, 1, 101-112.

Boekkooi-Timminga, E., & van der Linden, W.J. (1987). Algorithms for automated test construction. In F.J. Maarse, L.J.M. Mulder, W.P.B. Sjouw, & A.E. Akkerman, Computers in psychology: methods, instrumentation and psychodiagnostics. Lisse: Swets & Zeitlinger.

Dantzig, G. (1957). Discrete-variable extremum problems. Operations Research, 5, 266-277.

de Gruijter, D.N.M. (1986). Item banking with random or stratified tests. Tijdschrift voor Onderwijsresearch, 11, 61-66.

Ebel, R.L. (1967). The relation of item discrimination to test reliability. Journal of Educational Measurement, 4, 125-128.

Garfinkel, R.S., & Nemhauser, G.L. (1972). Integer programming. New York: Wiley.

Jensema, C (1976). A simple technique for estimating latent trait mental test parameters. Educational and Psychological Measurement, 36, 705-715.

Lord, F.M. (1980). Applications of item response theory to practical testing problems. Hillsdale, N.J.: Lawrence Erlbaum.

Lord, F.M., & Novick, M.R. (1968). Statistical theories of mental test scores. Reading: Mass.: Addison-Wesley.

Molenaar, W. (1974). De logistische en de normale kromme. Ned. Tijdschrift voor de Psychologie, 29, 415-420.

Rasch, G. (1980). Probabilistic models for some intelligence and attainment tests. Chicago, Ill.: The University of Chicago Press. (2nd edition).

Theunissen, T.J.J.M. (1985). Binary programming and test design. Psychometrika, 50, 411-420.

Theunissen, T.J.J.M. (1986). Optimization algorithms in test design. Applied Psychological Measurement, 10, 381-390.

Theunissen, T.J.J.M. & Verstralen, H.H.F.M. (1986). Algoritmen voor het samenstellen van toetsen [Algorithms for constructing tests]. In W.J. van der Linden (red.), Moderne methoden voor toetsconstructie en -gebruik. Lisse: Swets & Zeitlinger.

van der Linden, W.J. (Ed.) (1986a). Test item banking [Special issue]. Applied Psychological Measurement, 10 (4).

van der Linden, W.J. (1986b). Item banking met een dialoog gebaseerd op klassieke item- en testparameters [Item banking with a dialogue based on classical item and test parameters]. In G.R. Buning, T.J.H.M. Eggen, H. Kelderman & W.J. van der Linden (Ed.), Het gebruik van het Raschmodel voor een decentraal toetsservicesysteem (Rapport 86-3). Enschede: University of Twente, Department of Education.

van der Linden, W.J. (1987). Automated test construction using minimax programming. In W.J. van der Linden (Ed.), IRT-based test construction (Research Report 87-2). Enschede: University of Twente, Department of Education.

van der Linden, W.J., & Boekkooi-Timminga, E. (1988a). A zero-one programming approach to Gulliksen's matched random subtests method. Applied Psychological Measurement, 1, to appear.

van der Linden, W.J., & Boekkooi-Timminga, E. (1988b). A maximin model for test design with practical contstraints. Psychometrika, to appear.

Wagner, H.M. (1975). Principles of operations research. London: Prentice-Hall.

Table 1

Results for tests constructed from Item Bank 1

(Rasch model; n = 20)

| Maximal α as Objective | | | Maximal α as a Constraint | | |
|---|---|---|---|---|---|
| v | α* | α | p | α* | α |
| 5.0 | .8096 | .8478 | .30 | .6922 | .7866 |
| 4.5 | .8028 | .8413 | .40 | .7245 | .7956 |
| 4.0 | .7803 | .8252 | .50 | .7381 | .8004 |
| 3.5 | .7491 | .8069 | .60 | .7373 | .7997 |
| | | | .70 | .7136 | .7896 |
| | | | .80 | .6441 | .7559 |
| | | | .90 | .4131 | .6701 |

Table 2

Results for tests constructed from Item Bank 2

(Three-parameter model; n = 20)

| v | $\alpha^*$ | $\alpha$ |
|---|---|---|
| 5.0 | .8201 | .8579 |
| 4.5 | .8252 | .8607 |
| 4.0 | .8199 | .8551 |
| 3.5 | .8045 | .8460 |
| 3.0 | .7874 | .8386 |

Table 3

Results for tests constructed with Model (23), (25), and (26) and Model (37) – (39) from Item Bank 1 and 2

| | | Model (23), (25), (26) | | Model (37) – (39) | |
|---|---|---|---|---|---|
| Item Bank | n | $\alpha^*$ | $\alpha$ | $\alpha^*$ | $\alpha$ |
| 1 | 20 | .8096 | .8478 | .8107 | .8465 |
| 1 | 40 | .9013 | .9122 | .9020 | .9122 |
| 2 | 20 | .8201 | .8579 | .8256 | .8603 |
| 2 | 40 | .9074 | .9188 | .9096 | .9196 |

Titles of recent Research Reports from the Division of Educational Measurement and Data Analysis, University of Twente, Enschede, The Netherlands.

RR-87-1 R. Engelen, Semiparametric estimation in the Rasch model

RR-87-2 W.J. van der Linden (Ed.), IRT-based test construction

RR-87-3 R. Engelen, P. Thommassen, & W. Vervaat, Ignatov's theorem: A new and short proof

RR-87-4 E. van der Burg, & J. de Leeuw, Use of the multinomial jackknife and bootstrap in generalized nonlinear canonical correlation analysis

RR-87-5 H. Kelderman, Estimating a quasi-loglinear models for the Rasch table if the number of items is large

RR-87-6 R. Engelen, A review of different estimation procedures in the Rasch model

RR-87-7 D.L. Knol & J.M.F. ten Berge, Least-squares approximation of an improper by a proper correlation matrix using a semi-infinite convex program

RR-87-8 E. van der Burg & J. de Leeuw, Nonlinear canonical correlation analysis with k sets of variables

RR-87-9 W.J. van der Linden, Applications of decision theory to test-based decision making

RR-87-10 W.J. van der Linden & E. Boekkooi-Timminga, A maximin model for test design with practical constraints

RR-88-1 E. van der Burg & J. de Leeuw, Nonlinear Redundancy Analysis

RR-88-2 W.J. van der Linden & J.J. Adema, Algorithmic Test Design Using Classical Item Parameters

Research Reports can be obtained at costs from Bibliotheek, Department of Education, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands.

A publication by
the Department of Education
of the University of Twente
P.O. Box 217
7500 AE Enschede
The Netherlands